RAIS data at the Industrial Relations Section - User Guide

June 5, 2020

1 Introduction

RAIS (*Relação Anual de Informações Sociais*) is a matched employer-employee administrative dataset that contains the universe of formal employment contracts in Brazil.

The Department of Economics of Princeton University has signed a cooperation agreement with the Brazilian Labor Ministry that allows Princeton researchers to work with the RAIS data under approval and supervision of the Industrial Relation Section. The agreement includes provisions to secure the data at the IR Section. For additional information, please see the agreement itself.

This document is a summary note that gives an overview of the process of getting access to, and using the RAIS data at the IR Section server for research purposes.

Please be mindful that access to the data is based on trust: increased data security would require further restrictions to data use, which would be detrimental to research.

2 Data available

The anonymized RAIS data is stored in a secured IR Section server (*vm-amas*). The following datasets are available:

- RAIS Vínculos (1985-2017): dataset at the (job, year) level, with worker, establishment, firm and job characteristics.
- RAIS Estabelecimentos (1985-2017): dataset at the (establishment, year) level, which mostly aggregates information from RAIS Vínculos but also includes some extra variables (such as establishment opening date) and extra observations (establishments with no workers during the year who reported to the Ministry of Labor).

For additional information on the datasets, please refer to the data dictionaries and other documentation available.

Remark: deidentified cross-sections of the RAIS data are publicly available (all identifiers are removed), in case that is sufficient for your purposes.

3 Getting access to the Data at the IR Section

Eligibility: data access is restricted to Princeton University affiliated Researchers[1].

Authorization: data access requires authorization by the director of the Industrial Relations Section at the Economics Department, Alexandre Mas (amas@princeton.edu).

[1] Princeton University VPN access is required.

Statement of responsibility: the Brazilian Ministry of Labor requires researchers to sign a statement of responsibility (Annex II of the cooperation agreement, attached at the end of this document) and hand it in to the director of the Industrial Relations Section.

Code of conduct: the IR Section requires researchers to read and sign an internal code of conduct for the use of the RAIS data (see Appendix) and hand it in to the director of the Section.

IRB: consult the IRB to understand whether your project will require an IRB application[2].

Access to *vm-amas*: after authorization is granted, researchers will need access to the secured Linux server where the anonymized data is stored (*vm-amas*). IT staff of the IR Section will help with this step, as well as creation of a working directory. In case you need access to identified data, additional steps are necessary (see below).

4 Data use cases

4.1 Using the anonymized RAIS data in *vm-amas*

Most research projects do not require having access to actual individual or firm identifiers. In this case, researchers can use RAIS data with anonymized identifiers that preserve the panel structure of the data without revealing the identity of individuals. Anonymized RAIS data is accessible from one's computer, after establishing a VPN connection with Princeton University and accessing the secure server *vm-amas*.

If the research project uses *only* the anonymized RAIS data, it might not require IRB oversight. Please consult the IRB to clarify this point.

RAIS data: the anonymized RAIS data is available in the server *vm-amas*, folder */home/brdata/RAIS*. This folder is read-only.

Working directory: researchers will use their own project folders in *vm-amas* to store working datasets, do-files, etc.

Input: Uploading files into working directories will be unrestricted and trust-based (see attached Code of Conduct)

Output: Extraction of results will be trust-based (see attached Code of Conduct). You may download results, but **not data**, from *vm-amas*. However: before results are posted online, submitted for publication or conferences, they must be subject to screening by the director of the IR section to ensure they are compliant with the cooperation agreement between Princeton and the Brazilian Ministry of Labor.

4.2 Merging RAIS with external datasets

If the researcher needs to match the RAIS data to other external data sources, the project must be submitted to the IRB to obtain approval to manipulate crosswalks that contain private identifiable information. If approved, IR staff will temporarily authorize your access to the server *vm-brazil*, where original data are stored, along with the data necessary to anonymize it. The researcher will do the anonymization. In most cases, the researcher will only need the crosswalk of the original RAIS identifiers and the anonymized identifiers. In other cases, the researcher might require more information to anonymize the data, such as worker and firm names, or firm addresses.

After anonymization, the dataset will be placed in the the researcher's working directory in *vm-amas* and access to *vm-brazil* will be removed.

Steps for merging with external dataset:

- IR Section staff will temporarily grant access to *vm-brazil*, where original dataset and RAIS identifier crosswalks are available for the anonymization process. While the researcher has access to this server

[2]More information can be found in `https://ria.princeton.edu/human-research-protection/hrpp-home`

with identifiable information, he/she will not have access to *vm-amas*.

- Researcher anonymizes the dataset of interest, deleting from it any original identifiable information. The new anonymized dataset should be stored in a temporary folder in *vm-brazil*.
- Once anonymization is completed, IR Section staff will be contacted to copy the anonymized external dataset into the researcher's folder in *vm-amas*[3].
- Access to *vm-amas* will be restored. Researcher confirms that the dataset transfer to *vm-amas* has been successful and notifies IR Section staff. IR Section staff will then delete the temporary folder on *vm-brazil*.

For more details on the files with the original identifiers, as well as examples of how to merge the data, please see the Appendix.

4.3 Using RAIS data with original identifiable information

Occasionally, the research project hinges on the use of directly identifiable information. Some examples we could think of are: individual's names are used to identify kinship networks; establishment addresses are necessary for detailed spatial analysis; individual identifiers are used to conduct a telephone survey with a sample of workers in the data.

If it is the case, discuss it with the IR Section Director and staff to see which specific arrangements can be made.

5 Best Practice

5.1 Confidentiality

RAIS is a dataset containing sensitive information about workers and firms in Brazil. Please be mindful of the following best practice when using the data:

- Never search or try to identify specific individuals, establishments or firms while working with the data.
- Do not download disaggregate RAIS data. If this happens inadvertently, warn the IR Section staff and delete immediately.
- Read the cooperation agreement signed between the Department of Economics and the Brazilian Ministry of Labor, and comply with the rules therein.

5.2 Congestion

Researchers using RAIS data will be sharing server resources with other users. Here are some recommendations to ease congestion:

- Stata has a "*set niceness*" command[4] that allows to control the release of unused memory. The default is 5, but you can probably set it to higher levels.
- The virtual machine has limited storage space. Please limit unnecessary file duplication and clear up space after use.

[3] If the anonymized external dataset could be relevant for other researchers, please consider making it a public good by adding it to the */home/brdata* folder.

[4] `https://www.stata.com/manuals13/dmemory.pdf`

5.3 Public Good

RAIS data is a shared resource at the IR Section. It has been made available in .dta format, after a considerable time investment in getting access, cleaning and documenting the data. Here are some suggestions on how researchers using RAIS can contribute to the common good:

- If you find errors in the data, code, documentation, or other, please share with other users by adding a description of the problem in the shared folder */home/brdata/share*. You can also share recommendations for future iterations of the RAIS data, or even contribute with your own code (for instance, a particular way to correct a variable).
- If you match other datasets to RAIS that could be useful to others, please consider adding it to the server. This can be discussed with the IR Section.

Appendix A: Merging External Data

Data files with original identifiers, and crosswalks

- *id/rawtxt*: original *.txt* files received from the Brazilian Ministry of Labor. *RAIS Vínculos* files are in yearly folders *id/rawtxt/YYYY*, and *RAIS Estabelecimentos* files are in folder *id/rawtxt/estabelecimentos*.
- *id/YYYY* and *id/estabelecimentos*: for each *RAIS Vínculos* file received in year *YYYY* or each *RAIS Estabelecimentos* file, there is a corresponding zipped Stata *dta* file with the variables that have identifiable information (names, addresses, IDs, etc). Observations in each file have a variable *n* that indicates the observation number in the original file, so it can be merged back into the RAIS datasets if needed (see data dictionaries).
- *id/RAIS_employee_identifiers.dta*: crosswalk of variables *pis*, *pis_anonym*, *cpf* and *id_employee*. **Remarks**:
  - *pis* and *pis_anonym* are one-to-one (*pis_anonym* is just an anonymized version of *pis*)
  - each *pis* is associated to a unique *id_employee* and *cpf*, but the reverse is not true
  - not all *pis* have an associated *cpf* (e.g. because the *pis* is only observed in RAIS before 2002, when the data started containing workers' *cpf*)
  - each *cpf* is associated with a unique *id_employee*
  - for further information on how the correspondence between these variables is constructed, please refer to the do-files
- *id/RAIS_employerid_cnpjcei.dta*: crosswalk of *cnpjcei* and *cnpjcei_anonym* (establishment identifiers), which are one to one. Can be used both for *RAIS Vínculos* and for *RAIS Estabelecimentos*.
- *id/RAIS_employerid_cnpjraiz.dta*: crosswalk of *cnpjraiz* and *cnpjraiz_anonym* (firm identifiers), which are one to one. Can be used both for *RAIS Vínculos* and for *RAIS Estabelecimentos*.
- *id/sample1_RAIS_employee_identifiers.dta*: subset of *RAIS_employee_identifiers.dta* that corresponds to the 1% sample of *id_employee* in *RAIS/sample1.dta*.
- *id/sample01_RAIS_employee_identifiers.dta*: subset of *RAIS_employee_identifiers.dta* that corresponds to the 0.1% sample of *id_employee* in *RAIS/sample01.dta*.
- *id/cnpjsample1_RAIS_employerid_cnpjraiz.dta*: subset of *RAIS_employerid_cnpjraiz.dta* that corresponds to the 1% sample of *cnpjraiz* in *RAIS/cnpjsample1∗.dta*.
- *id/cnpjsample01_RAIS_employerid_cnpjraiz.dta*: subset of *RAIS_employerid_cnpjraiz.dta* that corresponds to the 0.1% sample of *cnpjraiz* in *RAIS/cnpjsample01∗.dta*.

Example 1: merging an external dataset of workers by *cpf*

Let's suppose we would like to merge RAIS Vínculos to a dataset *external.dta*. In this dataset, we have microdata on workers identified by their *cpf*. To allow the merge with the anonymized RAIS datasets, we need to anonymize *external.dta* using the same crosswalk. To do this, we merge *external.dta* to *RAIS_employee_identifiers.dta* using the *cpf* variable, keeping the variable *id_employee* as the anonymized identifier. Then we remove the *cpf*, as well as any other identifiable information, and the resulting dataset is now anonymized. The same idea can be applied to merge external datasets by *pis* and *CNPJ*, using the appropriate crosswalk (see above).

Example 2: merging an external dataset of workers by name

Suppose now that we would like to merge RAIS Vínculos to a dataset *external2.dta*, which contains microdata on workers identified by their names. An issue here is that there is no crosswalk ready between worker names (variable "*nome*" in RAIS) and anonymized identifiers such as *id_employee*. One possibility is to compile first a full list of worker names and *pis* from the files in folders *id/YYYY*[5]. Then we can use an algorithm to match names in this list with worker names in *external2.dta*. This will give us a correspondence between names in the external dataset and worker *pis*, which can be further matched to *id_employee* using the crosswalk *RAIS_employee_identifiers.dta* as in the previous example.

[5]Note that in these files, the data is not treated, so the same worker can have her name spelled in many different ways.

Appendix B: Code of Conduct

The Industrial Relations Section is making available for Princeton University researchers the Brazilian RAIS data. The data has been modified to protect confidentiality of workers and firms, in accordance with the agreement signed with the Brazilian Ministry of Labor.

The purpose of this Code of Conduct is to set the rules to ensure that all potential risks to the confidentiality of the data are considered and addressed. It is also meant to ensure that all individuals working with RAIS data are aware of their responsibilities when using it. In addition, the code has been written to meet legal requirements and best practice guidance.

1. **Purpose:**
    - The RAIS data shall be used with the sole purpose of conducting academic research.
2. **Integrity and Confidentiality:**
    - Our researchers shall **not** search or attempt to identify specific workers, establishments or firms in the RAIS data;
    - Our researchers shall read the user guide and comply with the rules therein.
3. **Security:**
    - Our researchers shall **not** download the RAIS data, full or partial, from the server. If it happens inadvertently, the downloaded data shall be deleted, and the IR Section staff shall be notified immediately;
    - Third parties without the authorization from the IR Section are **not** allowed to access the data;
    - Our researchers shall **not** distribute copies of the data to third parties.
4. **Generating Outputs:**
    - Graphs, codes, log files, regressions outputs and summary statistics can be extracted. These outputs shall **not** be attributable to individual workers, establishments or firms;
    - Outputs to be used for purposes **unrelated to publication** (e.g. lunch presentations at Princeton University, discussion with advisers, etc.) can be extracted without additional authorization;
    - Outputs to be used for purposes **related to publication** (e.g. posting a preliminary version online, submitting for conference or publication, etc.) have to be verified by the Director of the IR Section prior to publication.
5. **Cooperation Agreement with the Brazilian Ministry of Labor:**
    - Our researchers shall read and comply with the rules in the cooperation agreement signed by the Department of Economics and the Brazilian Ministry of Labor.

Appendix C: Statement of Responsibility (from Cooperation Agreement)

Annex II

STATEMENT OF RESPONSIBILITY

Name: _____

Date of birth: ____/_____/_____

Phone Number: (___) _____

Email: _____

Registration number: _____

I, a student identified above, duly enrolled in the _____, in view of the access to the Brazilian Annual Report of Social Information - RAIS, in a format of identified micro data, provided by the Ministry of Labor of Brazil, commit myself to:

I. Maintain confidentiality of the personal information contained in the abovementioned databases, to which I have access by virtue of my attributions, refraining from disclosing or publishing them, otherwise I will incur civil and criminal sanctions as a result of any improper use;

II. Handle the databases to meet specifically the object of interest of the Institution in which I study, as expressed in the signed Technical Cooperation Agreement;

III. Do not pass on the databases in an identified format;

IV. Maintain the necessary caution when displaying data on screen, printer, or even recording in electronic media, to prevent unauthorized persons from becoming aware of them;

V. Do not leave the terminal without closing the session of use of the bases, thus ensuring the impossibility of unauthorized access by unauthorized persons;

VI. Observe and comply with Good Practices on Information Security and its guidelines, according to Annex III.

I declare, on this date, to be aware of the responsibilities incumbent upon the use of the abovementioned database and to comply with the procedures described above.

[Place, date e signature]

[Two witnesses identified]